Fair boosting: a case study

Submitted for blind review

May 11, 2015

Abstract

We study the classical AdaBoost algorithm in the context of fairness. We use the Census Income Dataset (Lichman, 2013) as a case study. We empirically evaluate the bias and error of four variants of AdaBoost relative to an unmodified AdaBoost baseline, and study the trade-offs between reducing bias and maintaining low error. We further define a new notion of fairness and measure it for all of our methods. Our proposed method, modifying the hypothesis output by AdaBoost by "shifting the margin" for the protected group, outperforms the state of the art for the census dataset.

Although there are several papers on "fair" versions of learning algorithms such as naive Bayes, decision tree learning or logistic regression, boosting, which is one of the most successful and most widely used machine learning algorithms, has not been studied in the context of fair learning before. In addition to its popularity, boosting is an interesting framework in which to study fairness because notions such as a weak learner and the boosting margin have natural interpretations for fairness. We rigorously define these notions in Section 2 and analyze them in Section 3.

Following previous literature, we assume that the training data is biased against data points with a given feature value but we do not have access to the unbiased ground truth. We want to learn a classifier which has minimal error (as evaluated on the biased data) among all classifiers that achieve statistical parity. Dwork et al. (2012) point out that bias represents a notion of group fairness rather than individual fairness, and that it is still possible to discriminate against individuals even when achieving statistical parity. Thus, in addition to learning a classifier that has both low bias and error, we want a classifier that performs well on a measure of individual fairness. In this paper, we introduce a notion of fairness that captures how resistant a classifier is to uniform random noise on the training labels of the data points with a given feature value.

The Census Income Data Set (Lichman, 2013) is a widely used data set for machine learning research in which the learner's goal is to predict whether an individual's income exceeds \$50k per year based on census data such as age, education, gender, and marital status. In particular, when considering gender as a protected attribute, the dataset exhibits high bias. We use this data set as a case study to understand the fairness properties of the AdaBoost algorithm of Freund & Schapire (1997). We review the boosting algorithm and provide information about the Census data set in Section 1.

A primary advantage to using boosting is that boosting has a natural notion of margin which we can take advantage of to try to decrease bias while keeping error low. Our main empirical finding is that after boosting is performed to produce a hypothesis h, flipping the output label of h according to the boosting margin of the protected group outperforms the state of the art on the Census dataset both in terms of bias and label error. We compare this to data massaging (introduced by Kamiran & Calders (2009)), replacing a standard weak learner with a "fair" weak learner, and uniform random relabeling. Finally, in Section 4 we interpret and discuss our results.

1. Background

1.1. Notions of fairness

The study of fairness in machine learning is still young, and to the best of our knowledge we are the first to study the fairness properties of boosting. There are two prominent definitions of "fairness" that have been studied in the literature. The first is *discrimination* or *bias*, which for a distribution D over a set of labeled examples X with label $l : X \to \{-1, 1\}$ and a protected subset $S \subset X$ is defined as the difference in probability of an example in S having label 1 and the

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

probability of an example in S^C having label 1, i.e.

$$B(X, D, S) = \left| \Pr_{x \sim D|_{S^C}}[l(x) = 1] - \Pr_{x \sim D|_S}[l(x) = 1] \right|.$$

Similarly the bias of a hypothesis h is the same quantity with $h(x)$ replacing $l(x)$. If a hypothesis has low bias we say it achieves *statistical parity.* S represents the group we wish to protect from discrimination, and the bias represents the degree to which they have been discriminated against.

Dwork et al. (2012) point out that while bias is undesirable, it does not account for all possible forms of unfairness — it is a measure of group fairness rather than individual fairness.

The second notion, due to Dwork et al. (2012), measures individual fairness and requires a metric on the underlying set X. They call a learning algorithm "individually fair" if the output of the learning algorithm is similar for individuals which are close in X.

In this light, we define a new notion of fairness that departs from previous literature in that it does not require a metric on the underlying space. Rather, it makes the assumption that the process generating the bias is uniformly random, and measures the ability for an algorithm to recover the true labels from the biased dataset.

1.2. AdaBoost

Boosting algorithms work by combining *base hypotheses*, "rules of thumb" that are barely more accurate than random guessing, into highly accurate predictors. On each round, a boosting algorithm will change the weights of the data points and find the base hypothesis that achieves the smallest weighted error on the sample. It always increases the weights of the incorrectly classified examples, thus forcing the base learner to improve the classification of the examples that are the hardest to classify correctly. In this paper, we focus on AdaBoost, the most ubiquitous boosting algorithm. The algorithm is given in Algorithm 1. In all of our experiments we boost decision stumps for $T = 20$ rounds (after which accuracy does not significantly improve).

Given hypotheses h_i with weights α_i computed by AdaBoost, we call the *margin* of $x \in X$ is

$$\text{margin}(x) = \frac{\sum_i \alpha_i h_i(x)}{\sum_i \alpha_i}.$$

Note that this is different from the usual definition of the margin: in this paper, we do not multiply the combination of the base hypothesis by the correct label. The reason is that we want a measure of "confidence" of AdaBoost which can be computed by the learning algorithm even when it does not have access to the correct labels. Tt is well known that examples with small margins in magnitude contribute to most of the error of AdaBoost (Schapire et al., 1998). As such, flipping small-margin examples should maintain low error, and we show in our experiments that this is indeed the case.

Algorithm 1 AdaBoost (Freund & Schapire, 1997)

```
for $i = 1$ to $m$ do
  $D_1(i) = \frac{1}{m}$
end for
for $t = 1$ to $T$ do
  $h_t$ = base hypothesis with smallest error
  $\epsilon_t = \sum_{i=1}^{m} D_t(i)(1 - \delta_{h_t(x_i), y_i})$
  $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$
  $Z_t = 2\sqrt{\epsilon_t(1 - \epsilon_t)}$
  for $i = 1$ to $m$ do
    $D_{t+1}(i) = \frac{D_t(i) e^{-h_t(x_i) y_i}}{Z_t}$
  end for
end for
$g = \sum_{t=1}^{T} \alpha_t h_t$
$h = \text{sgn} \circ g$
return $h$
```

1.3. Baseline statistics about the Census dataset

The Census Income dataset, extracted from the 1994 Census database, contains demographic information about 48842 American adults. The prediction task is to determine whether a person earns over \$50K a year. 16192 of the people in the dataset are female, 32650 are male. 30.38% of men and 10.93% of women earn more than \$50K, therefore the bias of the dataset is 19.45%.

AdaBoost achieves an error of 15% after 20 rounds of boosting. The bias of the classifier output by vanilla AdaBoost is 18%. We note that simply removing the protected feature from the data does not reduce bias at all in this case since the classifier output by vanilla AdaBoost trained for 20 rounds on the full data doesn't (explicitly) use gender.

2. Methods

We define our methods. In what follows X is a labeled dataset, $l(x)$ are the ground truth labels, and $S \subset X$ is the protected group. We further assume that members of S are less likely than S^C to have label 1, which is true of the Census dataset when S is the women, for example. First we describe three rela-

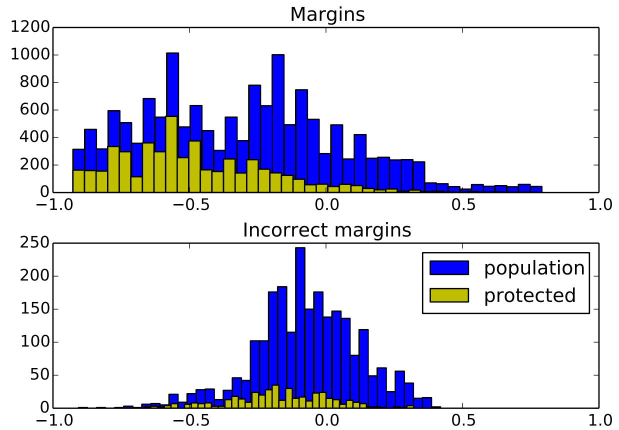

Figure 1. Histogram of boosting margins for the Census data set. The vast majority of women are classified as -1, and the incorrect classifications are closer to the decision boundary.

beling algorithms. A relabeling algorithm, when given a hypothesis h and a labeled data set X, l, produces a new hypothesis h' that uses h as a black box and flips the output of h according to some rule.

The *uniform random relabeling* (RR) algorithm computes the probability p for which, if members of S with label -1 under h are flipped by h' to $+1$, the bias of h' is zero in expectation. h' is then the randomized classifier that flips members of S with label -1 with probability p and otherwise is the same as h.

The *margin-shift relabeling* (MSR) algorithm computes the value θ such that bias is minimized by "shifting the margin" for examples from S from zero to θ. That is, $x \in S$ then $h'(x) = 1$ iff $\text{margin}(x) >= \theta$, and otherwise $h'(x) = \text{sign}(\text{margin}(x))$ as usual.

Next, we define *random massaging* (RM). Massaging strategies, introduced by Kamiran & Calders (2009), involve eliminating the bias of the training data by modifying the labels of data points, and then training a classifier on this data in the hope that the statistical parity of the training data will generalize to the test set as well. In our experiment, we massage the data randomly; i.e. we flip the labels of S from $+1$ to -1 independently at random to achieve statistical parity in expectation.

Finally, in *fair weak learning* (FWL) we replace a standard boosting weak learner with one which tries to minimize a linear combination of error and bias and run the resulting boosting algorithm unchanged. The weak learner we used computed the decision stump which minimizes the sum of label error and bias of its induced hypothesis.

To measure fairness, we test how these algorithms are resistant to uniform random noise that introduces bias against a random subset of the individuals. This is formalized as follows:

Definition 1. *We define the* uniform bias individual fairness *(UBIF) of a learning algorithm A on a labeled dataset X, l as follows. Introduce a new uniform random binary feature z on elements of X. Flip the labels of examples x that have $z = 0$ independently with probability p to -1 to get a new dataset X', l'. Run A on X', l' and let h be the resulting hypothesis. The uniform bias individual fairness of A is the fraction of flipped examples $x \in X'$ for which $h(x) = l(x)$.*

In our experiments we set $p = 0.2$. UBIF can be thought of as the following experiment: A learning algorithm is given a dataset in which bias has been generated uniformly at random. That is, we change the labels of a few individuals based on a feature which is blatantly random with respect to the classification task. For example, we purposefully flip a few labels in the data set of individuals who prefer chocolate ice cream over vanilla ice cream. The goal of an algorithm is to then recover the ground truth labels in the original dataset, recovering from the egregious bias against chocolate ice cream lovers. This models the ability of the algorithm to recover from bias against a few individuals.

3. Results

In this section we state our experimental results. They are summarized in Table 1. We also included the numbers for the Learning Fair Representations method of Zemel et al. (2013). In that paper, the authors implemented three other learning algorithms, these are unregularized logistic regression, Fair Naive-Bayes (Kamiran & Calders, 2009), and Regularized Logistic Regression (Kamishima et al., 2011). These methods all had errors above 20%; thus we see that our margin-based relabeling methods outperform the state of the art. To investigate the trade-offs made by these relabeling methods more closely, Figure 2 shows the rate at which error increases as bias goes to zero.

4. Discussion

Margin shift relabeling (MSR) performs equally to or outperforms every proposed method on all measures of bias and error, including the previous state of the art for the census dataset. Indeed, there is significant theoretical justification that shifting the decision boundary for the protected group achieves relatively high levels of fairness. While it is always possible to

	AdaBoost	RR	MSR	RM	FWL	LFR (Zemel et al., 2013)
label error	0.15	0.20	0.18	0.18	0.18	0.25
bias	0.18	0.00	0.00	0.03	0.07	0.00
UBIF	0.44	0.46	0.54	0.54	0.48	n/a

Table 1. A summary of our experimental results for relabeling, massaging, and the fair weak learner. The threshold for margin threshold relabeling (MTR) and the shift for margin shift relabeling (MSR) were chosen to make statistical parity zero.

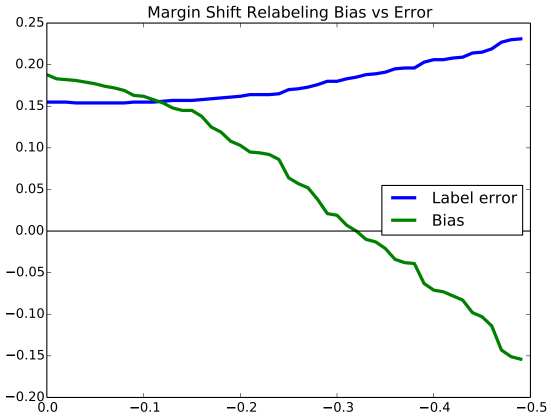

Figure 2. Trade-off between bias and error for margin shifting. The horizontal axis is the shift used for MSR.

shift the decision boundary until statistical parity is achieved, the risk is that some of the data points with changed labels are now labeled incorrectly, increasing error. For example, when the data points that are relabeled are chosen randomly, as in RR, each is now likely to be misclassified, resulting in an additional 5 percent error as seen in Table 1. To decrease error, we want to find the data points whose labels boosting was the most unsure of since these are more likely to be classified incorrectly by boosting. This means we should choose the data points to relabel with the smallest margin, as in MSR. Figure 1 shows that the distribution of margins for women is noticeably shifted when compared to the whole population, giving empirical evidence that this approach is sensible.

Of course, how data points with small margin are relabeled does matter. If it is done symmetrically so that both points labeled -1 and 1 are flipped, then it takes a larger threshold to achieve statistical parity when compared to MSR, which only flips labels from -1 to 1. This means fewer points need to be flipped, which in turn decreases error when compared to a symmetric version. It outperforms the baseline RR, where margin is not considered, showing that the points with small margin (in absolute value) are indeed less likely to be labeled correctly than points with large margin.

Replacing a standard weak learner by a weak learner that tries to minimize a combination of error and bias (FWL) does empirically reduce bias, but does not quite achieve statistical parity. Moreover, the label error of FWL is not better than that of MSR, and the trade-off between label error and bias cannot easily be controlled. The same is true for random massaging (RM).

A natural baseline for UBIF is 0.5, since a hypothesis chosen uniformly at random will flip back half of the points that were flipped to -1. Unmodified AdaBoost encodes the bias introduced, performing worse than 0.5. The question is whether we can recover from this randomly introduced bias, while still achieving low label error and low overall bias. Under RR, UBIF increases marginally, as it randomly flips labels back to a label of 1, a few of which were the points randomly biased against. MSR performs better than RR, indicating that the points that were randomly biased against have small margin under boosting.

A further advantage of MSR is that the trade-off between label error and bias can be controlled after training. To decide how much bias and error we want to allow, we do not have to fix the value of a hyperparameter before training the algorithm, unlike for most other fair learning methods. This means that the computational cost of choosing the best point on the trade-off curve is very low.

While these results are preliminary, they show the advantages of fair boosting: the margin can be used to find a superior classifier. We also give preliminary results that suggest the usefulness of measuring an algorithm's resistance to uniform bias.

References

Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM, 2012.

Freund, Yoav and Schapire, Robert E. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. Syst. Sci.*, 55 (1):119–139, 1997.

Kamiran, Faisal and Calders, Toon. Classifying without discriminating. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pp. 1–6. IEEE, 2009.

Kamishima, Toshihiro, Akaho, Shotaro, and Sakuma, Jun. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 643–650. IEEE, 2011.

Lichman, M. UCI machine learning repository, census income data set, 2013. URL `http://archive.ics.uci.edu/ml`.

Schapire, Robert E, Freund, Yoav, Bartlett, Peter, and Lee, Wee Sun. Boosting the margin: A new explanation for the effectiveness of voting methods. *Annals of statistics*, pp. 1651–1686, 1998.

Zemel, Rich, Wu, Yu, Swersky, Kevin, Pitassi, Toni, and Dwork, Cynthia. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pp. 325–333, 2013.